Multimodal Access to Georeferenced Mobile Video through Shape, Speed and Time

Sérgio Serra
LaSIGE, Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal
sergioserra99@gmail.com

Ana Jorge
LaSIGE, Faculdade de Ciências
Universidade de Lisboa
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
ananunesjorge@gmail.com

Teresa Chambel
LaSIGE, Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal
tc@di.fc.ul.pt

Video is becoming increasingly accessed, captured, and published on the Web, from different platforms and devices. Users can easily georeference the information they capture and access, allowing to enrich their contextualization. But video search has been limited to keywords, or a set of parameters, providing limited support for temporal and spatial dimensions. We propose novel ways to search and access georeferenced videos, where these dimensions are of central importance, especially by video trajectories shape and speed, and by time, using a multimodal interactive mobile interface, involving gestures and movement, with the potential for more natural interactions, increased engagement, sense of presence and immersion. The evaluation based on high-fidelity prototypes had positive results. Users found most features useful, satisfactory, sometimes fun, and easy to use. Different options and modalities were found interesting and adequate for different use scenarios that could be identified and suggested, and some concerns and challenges were identified to be taken into account in future developments, towards more flexible and effective interactive content access, through more natural interaction with mobile devices on their own or as second screens to a larger screen on TV or public displays.

Georeferenced videos; Gestures; Multimodal; Search; Browsing; Consumption; Space; Speed; Shape; Time; Movement; Trajectories; Mobile; Second Screen

1. INTRODUCTION

A large amount of digital video is being uploaded everyday to the Web. It is becoming widely captured, shared and accessed from different platforms and devices, and increasingly it can be georeferenced, allowing to enrich its contextualization. Although a lot of videos are available to search and watch, the current and most used mechanisms to browse and find them are based on a limited set of parameters such as: keywords, duration, video quality, ignoring the temporal and spatial dimensions. Video has an enormous potential for immersion and mobile devices allow to access information while 'immersed' in reality anywhere. With the proliferation of devices like: smartphones, tablets and more recently wearables, we could take advantage of the multimodal sensors available, to create new ways to find and navigate georeferenced videos through space and time (both along the video timeline and when they were captured) and space, using more natural interfaces, involving gestures and movement shape and speed, with the potential for increased engagement, sense of presence and immersion when accessing the videos.

This work builds on previous work done in the context of Sight Surfers (Noronha et al. 2012) (Ramalho & Chambel 2013a, 2013b), an interactive web application for sharing, visualizing and navigating georeferenced 360° interactive videos, as hypervideos, including city tours or more extreme activities like kart racing. These can be experienced

in increased immersion and isolation, or synchronized with a map while being played. To provide a more complete support for the additional spatio-temporal dimension in georeferenced videos, and keeping the purpose of increasing immersion, aligned with the augmented sensorial experience, we want to create richer mechanisms for interactive search, visualization and navigation in more natural modes of interaction.

In this paper, we describe our work in this direction. The next section highlights main challenges and opportunities, and presents most relevant related work. Next, the conceptual model and design options are presented for the multimodal georeferenced mobile video access in space and time, demonstrated in the prototypes and evaluated in the following section. A user evaluation was conducted with a high fidelity prototype, to find out about perceived usability and acceptance, focusing on usefulness, satisfaction and ease of use. Finally, the paper ends with conclusions and perspectives for future work.

2. RELATED WORK

Challenges for this work include providing users with an adequate interactive interface capable of capturing and expressing the temporal and spatial dimensions, allowing to represent speed and trajectories shapes and speed along time, while offering an intuitive, simple, effective and natural way to search for, and present resulting videos and

navigate them in a mobile environment. It is both a challenge and an opportunity because users are not used to searching and navigating in these dimensions, but technology is allowing to capture movement in mobile devices in ways that hold the potential to support more natural interactions involving time, shape and speed towards more immersive experiences.

Most video libraries and websites like YouTube or Vimeo are based on keywords and have at most a very limited support to access video based on spatial and temporal dimensions. Rego et al. (2007) developed VideoLIB, a digital library that enhances video retrieval by using spatial and temporal operators, based on Dublin Core and MPEG-7 meta-data standards. Search criteria include action (what), person (who), time (when) and place (where), and use operators like before, during and after to define time intervals. This allows to make searches like “retrieve Madonnas’s video clips which were produced outside the USA during 1990’s”. It uses a form and text based interface, without the use of maps, and videos are considered as a whole - trajectories and speed are not taken into account.

There are some approaches to search and browse videos, and mainly photos, using maps. Google Street View is a 360° photo viewer using a spherical image projection and geolocalization, but it does not provide video, nor user generated and alternative views of the places. Panoramio (.com) is a georeferenced photo sharing website accessed as a layer in Google Earth and Google Maps. Users can do text-based search or navigate in the maps, and view photos taken by other users, based on location. The photos are presented along with a map that highlights their location, both as a collection resulting from a query or one by one. There are filters to highlight most popular, recent, famous places and indoor, both on a separate tab with the filtered photos, and by enlarging these photos among those shown on the map. Finsterwald et al. (2012) developed The Movie Mashup Application (MOMA), as a public web map-based service for searching movies based on location, combining geotagged resources and text processing, mashing up information from DBpedia, GeoNames and Wikipedia synopses. Through its GUI, it allows to search and browse a data set of movies by director, location in text, by polygonal areas in the map, from locations extracted from movie titles, to compare query distributions and, using a mobile device version, allows to query for movies whose action took place around the user’s current location. Although maps are a natural way to represent georeferenced information, and video often involves a trajectory, most solutions only allow users to post or access videos based on a single GPS location (usually the initial position). Hao et al. (2011) present user-generated videos that relate to geographic areas in a map interface. They focus on the automatic selection of keyframes

to represent the videos, and the determination of the location to place them on the maps. So they emphasize hotspots that are shot in the videos in front of the shooting spot, and not so much on their trajectories.

Concerning spatial and haptic interactive search in a mobile environment, in the last years we noticed a growing popularity of gesture interfaces and second screen applications. Lei & Coulton (2009) implemented a gesture controlled application that act as a wand, using mobile sensors. It allows both proximity and remote search of points of interest (POIs) based on the orientation of the wand, as an interactive spatial ‘Flashlight’, and the possibility of users to create additional content for a particular POI as photographs tagged with POI’s location and the direction from which the photograph was taken. Photographs can then be filtered based on a desired viewing angle in a real world environment. Premraj et al. (2010) presented iWalk, a tool that allows multimedia exploration of geo-tagged data through movement, to move through the digital space of a collection, and gesture, for direct data manipulation (e.g. select, go to next, zoom). They experimented with geo-tagged photographs and sound collections, and a non geo-tagged museum collection, where the user defined a mapping between digital and physical spaces. Their approach makes use of computer vision algorithms computed on standard commercial camera inputs and is able to operate in real-time.

Mobile devices may also act as second screens (Courtois 2012) to complement and interact with larger screens like TV or even public displays. MobiToss, an application created by Scheible et al (2008), allows for mobile multimedia art sharing and creation. By using a mobile device with built-in accelerometer sensors, users can take a photo or video and “throw” it onto a large public display, with a gesture, for viewing and manipulation, through tilting. The users-created clips are augmented by the system with items like music or brand names and sent back to their phones as personal artifacts of the event. The preliminary user evaluation showed that capturing and throwing mobile content onto a large screen and manipulating it with gesture control into an art piece was perceived as an intuitive and fun activity. They enjoyed and engaged in the experience and appreciated getting something out of it, especially something artistic. But it requires improvements, by applying a more balanced set of video effects, adding group interaction and a more intuitive UI, to accommodate different movements users do to throw, and increase the perception of what is going on. This work explores natural gestures with a mobile device to manipulate photos or videos as a second screen, but in doing so, it does not explore spatial and temporal dimensions in videos. And none of the related work found addresses speed and trajectories in video as we propose to do.

Figure 1. Different inputs, through: a) Gesture (speed); b) Touch speed; c) Travelling, e.g. in a car; d) Gesture (shape); e) Touch for shape in a map; and f) In time for older videos.

3. VIDEO ACCESS BY SPACE, SPEED AND TIME

Space and time dimensions are taken into account primarily in videos' locations, trajectories shapes and speed, and the time inside and about the video. To explore the interactive interfaces and user experience with these dimensions in georeferenced videos, a high fidelity prototype was developed for Android, and variations were designed for different use scenarios, exploring the use of sensors and location services, and following our preliminary work with low-fidelity prototypes (Serra & Chambel 2013). Next, we present the rationale behind main design options.

3.1 Through Touch – with finger

This is the most conventional interface, that allows to draw the query shapes by touching the screen (Fig.1d), or even on a touch pad of a laptop or desktop. The speed of this drawing may also be captured to query by speed only, or by both shape and speed. This kind of interaction may be more familiar and provide for better accuracy than the following, especially when the user has a hand free for this interaction

3.2 Through Gesture – with mobile

This modality can be used by moving the mobile in a gesture, to draw a shape (Fig.1d) or demonstrate a speed level (Fig.1b). The gesture speed is captured using the accelerometer, but this sensor did not allow to capture the shape with enough precision. So we

Figure 2. Different outputs: a) List with colored timeline; b) List with grey timeline; c) Results on a Map showing videos trajectories d) List with one-color timeline.

also use the gyroscope for shape capture, and the movement now is based on leaning the phone around to draw the shape, by letting the cursor slide down in accordance to the inclination (Fig.1d).

This can be done with the hand that is holding the device, even if the other one is not free, and has the potential for a more natural or immersive modality to select the desired videos, to watch on the mobile or on a wider screen like a TV. In this context, viewers are used to interacting with a control in one hand, and keeping focused on the video on the wide screen.

3.3 Through Traveling – on the move

When on the move, users are often travelling by car, train, subway, planes or even walking or running. In addition or as an alternative to use your current location to access videos shot in the same location, it can be interesting to take the chance, especially when not driving, to watch videos that were shot at a similar speed, and be able to enjoy the viewing experience in a more immersive way, by matching the speed of what you are seeing with the speed that you are feeling or experiencing in reality (Fig.1c). This might have a special impact in high speed videos in more extreme activities, but one could e.g. search for a video of the current walking trajectory at a similar speed some years ago, to compare what could be experienced back then in the same location. As in the previous cases, both speed and trajectory shape can be captured this way. Speed and even location might have more potential for immersion in the

immediate video viewing, to get related videos on the spot. But capturing a trajectory can also be interesting, to search for other videos that have similar paths even if in different places, e.g. another similar Kart race elsewhere in the world. While gestures rely on sensors, travel features rely on location services like GPS.

3.4. Where - on the Map or Current Location

Since videos on our X application are georeferenced, search can be made dependent on their location. Users can use a map to select locations (Fig.1c), or the current location can be captured, for the queries (besides the possibility to specify a set of places, like cities, as in more traditional interfaces).

3.5 Anywhere

Videos may also be searched independent of their location. In this way, a map is not used and only speed and shape are drawn on the screen or in the air, or captured on travel without a geo-reference.

3.6 Results in Maps or Lists

The resulting videos can be presented as trajectories on a map (Fig.2c), where each trajectory can also be seen as the video timeline synchronized with the video as in Sight Surfers (Noronha et al. 2012). And this is the default when search is based on a location. But results may also be presented independent of their location, e.g. in a list, where the speed (Fig. 2abd) and or the shape can be emphasized in each video timeline. Also note that users can switch between map and list views and select what to show, for the same results. Search results are presented to the user in different design alternatives, each one offering different visual cues and information about the content retrieved, in terms of shape, speed and time, both in maps or lists.

3.6.1 Speed Awareness

Speed can vary along a video, so the results would present first the videos that keep the desired speed (with a tolerance) for longer, but still, it provides awareness of the segments where the speed is as queried for, higher and lower. Fig.2abd shows three alternative designs for presenting speed in the video *timelines* in: a) Color: green for the searched speed, red for faster and blue for slower; b) Gray-Scale: mid tone for searched speed, darker for faster and lighter for slower; and d) Color Highlight: green for searched speed and two gray tones for faster and slower as in b), allowing for higher contrast of the searched for speed.

3.6.2 Shape Awareness

Shape is shown by default when on a map, but can also be presented in the list view, as in Fig.2c where each video timeline takes the shape of the corresponding trajectory. Whereas speed awareness is optional in all the timelines: map, and list with or without shape.

3.6.3 Time Awareness and Search

Besides showing the video duration next to the video, and the current time on the timeline, the time when the video was shot is represented by color, either on a timeline on the lists (Fig.2abd) or on the video trajectory in the maps (Fig.2c), to show its age. We opted to loose saturation for older videos, while more recent ones keep a more vivid color.

On top of the screen, a colored timeline represents a time range corresponding to the “age” of the videos being shown. The oldest time on the left, with the lowest saturated color that becomes increasingly saturated towards the right. The time labels (a year in the example of Fig.2) for the time range are presented in text boxes to the left and right of the bar. This timeline makes the mapping of time and color explicit to the users, allowing them to identify newer and older videos in the time range. This timeline may have one colored bar, when speed is not shown (e.g. only green in the map of Fig.2c), or have three bars, one for each speed color (e.g. in gray tones or RGB in the list views of Figs.2ba) to emphasize how the different colors “age”.

In this view, users may search or filter videos by time, either by selecting the two textboxes on each top corner through touch, and writing the wanted year on a virtual keyboard, or using gestures to increase or decrease the value, by drawing a spiral shape with the finger on the screen (the drawn gesture highlighted in Fig1.d). A clockwise spiral shape means going forward in time, while the counter clockwise means going backward. This mapping was hypothesized to be the natural one considering the metaphor of the clock.

4. USER EVALUATION

We conducted a user study to evaluate the features designed and to investigate about preferred alternatives and users’ perception about usability and user experience, and their application in real use scenarios.

4.1 Method

We performed a task-oriented evaluation based mainly on Observation and semi-structured Interviews, after explaining the purpose of the evaluation and the concept behind the Sight Surfers application context, and the features being evaluated in the high-fidelity prototype. The order of the variants evaluated changed for each user following a within subject method. At the end of each task, users provided a 1-5 USE (Usefulness, Satisfaction, and Ease of use) rating (Lund 2001) about the tested interactive features, and were encouraged to make comments and suggestions. The evaluations took place in our university campus mostly in-doors where they could sit together with the evaluator and walk around.

4.2 Participants

There were 9 participants aged 18-27 (22,5 on average, 3 F, 6 M). All users had at least finished high school, 3 from computer science, the rest from a mix of backgrounds, all had a smart phone used on a daily basis to access info, and 9 often search for and watch videos but mainly on PCs, sometimes on tablets and seldom on mobiles.

4.3 Results

Main results are summarized by mean values for USE and most significant comments, for each of the categories of features.

4.3.1 Search by speed

Users were asked to “search for videos by speed, using touch and gesture”. Through touch: was considered quite fun and very easy to use, and found useful by most of the users (U:3.67;S:3.78;E:4.44), e.g. comment: *“Georeferenced search is without a doubt useful in many situations, for example to know parts of a city”, “I can imagine an athlete using this functionality to search for running tracks”*. Through gesture: had slightly lower scores and especially in ease of use (U: 3.22; S: 3.22; E: 3.67), in this new unfamiliar modality. Finally, users were asked to “search for videos using travel speed”. We asked them to “search for a video with speed corresponding to a fast paced walk” (U: 3.43; S:3; E:4.11). Some users pointed out that this feature could be more useful when traveling on a car or train, as they felt a bit awkward walking fast for the experiment and they were not then relaxed to watch the video at the same speed.

4.3.2 Search by shape

Users were asked to “search for a given shape in a specific location on a map through touch (using the finger)”. This feature received positive feedback (U: 3.44;S:3.11;E:3.44): Participants appreciated having the ability to georeference the shape, finding several use scenarios where it could be used, e.g.: *“As an athlete, I could use this search to find running tracks with certain shapes, and see them on a map”, “Georeferenced search is without a doubt useful in many situations, for example to know parts of a city”*. “Searching for videos with a ‘u’ shape by using the phone to draw” was found very difficult and not useful, receiving the most negative feedback from most of the participants (U: 1.89; S: 2.22; E: 2.22), *“I don’t think it’s practicable to use this while I’m doing other activities, like walking or even talking with other person.”* The balance between speed and sense of control is in our opinion far from optimal in the current state of the prototype, not allowing us to create the type of experience we had in mind. Also, users are not used to using the phone for this kind of action, tending to prefer the finger to do it.

4.3.3 Results in Maps or Lists

In general, users preferred the more familiar way of showing the videos in a **list view** (U:4;S:3.56; E:4.44), by being easy and simple, *“It’s the fastest way to see videos if the geographic localization doesn’t matter”*. Although the **map** received positive feedback too (U:3.44;S:3.11;E:3.78), and most said it was useful when looking for georeferenced trajectories, videos in specific places, to be aware of the video locations and trajectories length. Also: *“It’s nice to see the video trajectories [on the map] without the need to see the video”*. Main concerns referred to awareness of the amount of videos retrieved and the representation if in the presence of a huge amount. Filtering of results was not in the scope of this test, but it is aligned with our own concerns, being addressed (Ramalho & Chambel 2013a, 2013b). There were suggestions to have a mix of list and map, where while navigating a list of videos we could see these videos’ locations on a map, and when hovering on a trajectory, we could see more detailed information about the videos on that trajectory in a list (with speed, duration, video image, etc.).

Users found the **timelines** useful, satisfactory and an easy way to find the searched speed on the videos in the results. We presented three alternative designs, and the majority of the participants selected the one-color alternative as the preferred one (table 1). Most users preferred this color highlight (U:4.11; S:4.11; E:4.44) in the searched speed with the other speeds made less noticeable in grey for being easier to use, and then the color version (3.44;3.44;3.6), which they found useful and quite satisfactory although more difficult to use: *“The one-color timeline is an easy and direct way to find the right speed, the colored timeline with 3 colors is harder to understand and makes the interface look less clear, more busy and flashy”*. The grey version was found more difficult to use and less useful, satisfactory and even fun (U:2.56; S:2.44; E:2.67). These preferences for the alternative designs for speed awareness matched our expectations hypothesized in the design.

Table 1: Speed awareness in trajectories timelines

designs	Usefulness		Satisfaction		Ease of Use	
	M	σ	M	σ	M	σ
1 Color	4,11	0,31	4,11	0,5	4,44	0,83
Gray	2,56	0,68	2,44	0,5	2,67	0,67
Colored	3,44	0,76	3,44	0,68	3,44	0,88

After a search, users were asked to filter the results by “going backwards in time for older videos”, using the virtual keyboard to type the year, and gestures through touch. Results showed that users liked both approaches to filter the results by time, however they found the keyboard slightly more useful and easier to use (U:3.67; S:3.22; E:4.0), while the

gesture was more satisfactory and funnier (U:3.33; S:3.67; E:3.44). In a general perspective, users found that both approaches complement each other, the gesture being faster and the timeline more precise.

5. CONCLUSIONS AND FUTURE WORK

This paper presented the motivation and design options for georeferenced mobile video access in space and time, in a context of user generated content, built into a high-fidelity prototype that was developed in Android and went through a user evaluation to learn about user satisfaction and preferences about its features. Our focus relied on three main dimensions: shape, speed and time taking leverage of mobile multimodality to explore new ways to access and navigate georeferenced videos based on touch, gestures and movement.

The user evaluation was encouraging, showing that users found most features quite satisfactory, even fun, and easy to use, and different options and modalities were found interesting and adequate for different use scenarios that could be identified. Usefulness and ease of use were more readily associated with more familiar modalities, while natural gestures and movement were considered quite satisfactory or even fun sometimes. In particular, the gestures to navigate in time, through older and newer videos, were very well received by the participants, and were pointed out as an effective and fast way to search in time, that could combine well with a more precise, versatile and traditional way to introduce a specific time on a keyboard. On the other end, the least appreciated feature was the search by shape that for technical reasons was relying on an implementation that did not allow for as much control and precision as was designed in low fidelity. It adopted the mobile phone tilting instead of allowing to perform the gesture with the mobile in the user's hand that correspond to the intended shape, and would also allow for a better capture of the speed in the movement. This was considered a limitation in the current prototype beforehand, and it was reflected in the evaluation results. Another situation that can be improved is the conditions for the evaluation in more realistic situations where users may feel the need or benefits of using trajectories and speed for video access in the different modalities in real life, like those some of them identified in the comments, but that most of them are not used to having.

The next steps include refining and extending some of the current solutions based on feedback received by the users, and exploring new ways to navigate, e.g. in time, also during immersive video playback. Also second-screen is an area we are addressing, and some work is already being conducted to be able e.g. to "throw" the movie to a bigger screen, like a TV, to watch the video in a more immersive setting, while using the mobile device for navigational aid and access to related content.

6. ACKNOWLEDGMENTS

This work is partially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014 and the ImTV research project ref. UTA-Est/MAI/0010/ 2009.

7. REFERENCES

- Courtois, C., D'heer, E. (2012) Second screen applications and tablet users: constellation, awareness, experience, and interest. *Proc. EuroITV'12*, ACM Press, 153-156.
- Finsterwald, J. M., Grefenstette, G., Law-To, J., Bouchard, H., and Mezaour, A.D (2012) The Movie Mashup Application MoMa: Geolocating and Finding Movies. *Proc. GeoMM'12*, ACM, 15-18.
- Hao, J., Wang, G., Seo, B., and Zimmermann, R. (2011) Keyframe Presentation for Browsing of User-generated Videos on Map Interfaces. *Proc. ACM MM'11*, ACM Press, 1013-1016.
- Lei, Z., and Coulton, P. A. (2009) Mobile Geo-wand Enabling Gesture Based POI Search an User Generated Directional POI Photography. *Proc. ACE'09*, Int. Conf. on Advances in Computer Entertainment Technology, ACM, 392-395.
- Lund, A.M. (2001) Measuring usability with the USE questionnaire. *Usability and User Experience*, 8(2).
- Norman, D. (1993) *Things that Make us Smart*. Addison Wesley Publishing Company.
- Noronha, G., Álvares, C., and Chambel, T. (2012) Sight Surfers: 360° Videos and Maps Navigation. *Proc. GeoMM'12*, ACM Press, 19-22.
- Premraj, V., Schedel, M., and Berg, T.L. (2010) iWalk, A Tool for Interacting with Geo-Located Data Through Movement and Gesture. *Proc. ACM MM'10*, ACM Press, 1059-1062.
- Ramalho, J. and Chambel, T. (2013a) Windy Sight Surfers: sensing and awareness of 360° immersive videos on the move. In *Proc. of EuroITV 2013*, Como, Italy. 107-116.
- Ramalho, J., and Chambel, T. (2013b) Immersive 360° Mobile Video with an Emotional Perspective. *Proc. ImmersiveMe'2013*, ACM Press, 35-40.
- Rego, A., Baptista, C., Silva, E., Schiel, U., and Figueirêdo, H. (2007) VideoLib: a Video Digital Library with Support to Spatial and Temporal Dimensions. *Proc. SAC'07*, ACM, 1074-1078.
- Scheible, J., Ojala, T., and Coulton, P. (2008) MobiToss: A novel gesture based interface for creating and sharing mobile multimedia art on large public displays. *Proc. ACM MM'08*, ACM Press, 957-960.
- Serra, S., and Chambel, T. (2014). Towards Multimodal Consumption of Georeferenced Mobile Video Using Shape and Speed. In *Proc. of 2nd Int. Workshop on Interactive Content Consumption at ACM TVX 2014*.